Georgia Tech Research Corporation

Georgia Institute of Technology
Office of Sponsored Programs
Atlanta, Georgia 30332-0420 U.S.A.

February 13, 2007
In reply refer to: E-21-6F2

Dr. Harold Cheyne
Sensimetrics Corporation
Suite 305
48 Grove Street
Somerville, MA 02144-2500

Subject: Final Report
Project Director(s): Clements, Mark
Telephone No.: 404-894-4584
Contract No.: AGR SIGNED
Prime No: N/A
"Identification and Application of Acoustics cues of Vocal Effort Changes..
Period Covered: N/A

The subject report is forwarded in conformance with the contract/grant specifications.

Should you have any questions or comments regarding this report(s), please contact the Project Director.

Sincerely,

Kamie Cunningham
Date Entry Specialist

Addressee: 1 copy

Speech Modification for Distance Cueing

Kaustubh Kalgaonkar Mark Clements

Abstract

Current virtual audio environments can effectively simulate the direction to a sound source relative to the listener, but have difficulty simulating distance. The use of distance cues such, as overall energy level and direct–to–reverberant ratio, requires the listener to have prior knowledge of the intensity of the source and/or the reverberance of the acoustic environment. Listeners know from experience that a talker adjusts his/her vocal effort to compensate for the acoustic loss due to the distance separating them. A listener can use the learned acoustic correlates of vocal effort, together with the received sound level, to estimate distance to the talker. This work lists some of the acoustic parameters of the speech that change with distance and some techniques for artificially creating vocal effort changes in speech.

I. Introduction

Studies of vocal effort have identified several acoustic parameters that either are manipulated by a talker in response to an increase or decrease in talker-to-listener distance, or are effects directly related to the acoustic environment. Speech intensity and fundamental frequency (f0) have been analyzed in recordings of real speech at varying vocal efforts and talker–to–listener distances. Other parameters such as formant frequencies and amplitudes, or direct–to–reverberant ratio, have been quantified for changes in talker–to–listener distances. Some work has studied the effect of individual acoustic parameters such as intensity [1] or pitch (f0) [2] on the perceived distance of short phrases. Other work has used isolated vowels to study the effect of varied vocal effort [3] or varied talker–to–listener distance [4], upon perceived talker–to–listener distance and spectral changes with that distance, respectively. Direct–to–reverberant ratio and its effect on distance perception has been investigated [5]. While this work has suggested possible first steps in synthesizing or processing speech to manipulate the perceived talker–to–listener distance, no systematic study has been conducted to determine the most salient acoustic cues for vocal effort related

Authors are with Center for Signal and Image Processing, School of Electric and Computer Engineering, Georgia Institute of Technology. Atlanta GA 30332. Email: {kaustubh, clements}@ece.gatech.edu

to talker–to–listener distance. This paper details a systematic study of acoustic parameters related to talker–to–listener distance perception and sugges an algorithm for speech modification of vocal effort.

II. Recording of Stimuli and Acoustic Correlates

To identify the acoustic parameters that the talkers vary with changes in distance from the listener, reference recordings were made using real talkers and listeners in a large open field with various talker–to–listener distances. This open environment guaranteed low reverberation. Four (2 male, 2 female) native English speakers served as participants in the study. In all the experiments the listeners's position was fixed and the talkers's position was varied along 1, 1.4, 2, 2.8, 4, 5.6, 8, 11.2, 16, 22.4 and 32 meters. At each distance the talker was instructed to speak a preamble of several sentences as he/she would naturally speak to someone at that distance. The listener provided qualitative feedback to rate the naturalness of the speech. The speaker made adjustments based on this feedback. Following this the talker was asked to maintain the same degree of vocal effort and speak a series of words and sentences. The words were taken from those used in the Diagnostic Rhyme Test (DRT), a speech intelligibility test that uses pairs of monosyllabic words with minimally contrasting initial consonants.

A. Analysis

The voiced frame with the maximum energy was selected for analysis which, monitored, pitch, energy of the frame, formant frequencies, formant amplitudes, glottal response and spectral tilt as parameters. It was observed that the overall energy of the speech signal increases with increase in talker–to–listener distance. For both male and female speakers, overall energy of the utterance increases by 15dB over the range of 1 to 32m. The rate of increase in energy is higher over shorter distances. Figure 1 shows trend in the overall energy of utterance with respect to various talker–to–listener distances for both male and female talkers. The energy in the frame of maximum voicing also shows a trend similar to overall energy with respect to distance, suggesting that the energy increase is due to the signal energy rather than the signal duration.

Fundamental frequency shows a steady increase with distance. Overall increase in pitch over 1–32m was found to be 65 to 75Hz for male speakers and about 40 to 50Hz for female speakers. Figure 2 reflects this trend.

Formant frequencies do not show any consistent change with respect to distance, the amplitude of the second (A2), third (A3) and fourth (A4) formant on the other hand increase with talker–to–listener

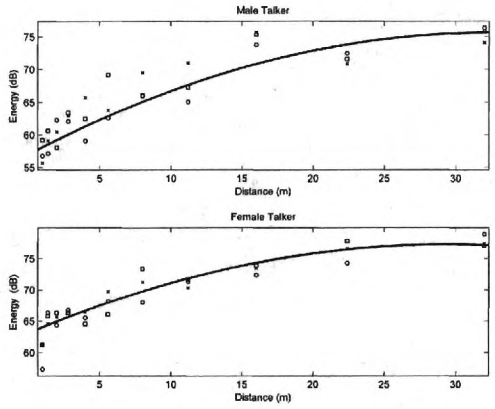


Fig. 1. *Energy Variation with Distance*

distance. The rate of increase of the third and the fourth formant amplitude is higher than that of the second formant. To capture this relative change in the formant amplitudes spectral tilt (ST) as defined in Equation 1, was used. Spectral tilt increased by 2dB over the distance of 16m. Figure 3 shows plots of the LP spectrum of voiced frames with maximum energy for talker–to–listener distance of 1m and 32m respectively, Note that the amplitude of second, third and the fourth formants have increased with talker–to–listener distance, with little or no change in frequency.

$$ST = \frac{A3 - A2}{ln_2 F3 - ln_2 F2} \tag{1}$$

Open Quotient, the ratio of time, the vocal folds are open during one fundamental period to the entire fundamental period did not show a consistent trend of change with respect to variation in talker–to–listener distance.

Based on the above observations in order to be able to modify the vocal effort of an arbitrarily recorded speech signal to match the vocal effort for a given listener–to–talker distance the pitch, formant amplitude and the energy of the utterance have to be modified concurrently. The pitch and the energy of the signal can be easily modified using the parametric LPC vocoder but it is not easy to modify the spectral tilt and formant amplitude directly using the LPC polynomial. The next section presents a computationally efficient method for modifying the formant amplitudes.

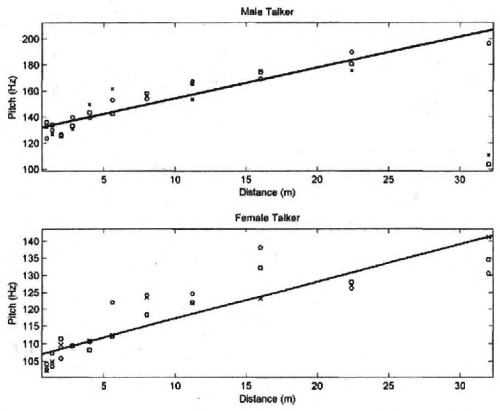


Fig. 2. *Pitch Variation with Distance*

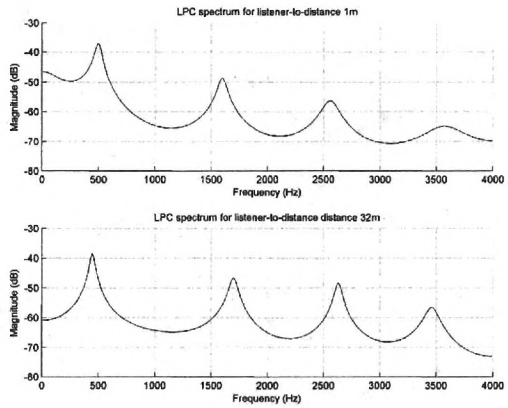


Fig. 3. *Formant Variation with Distance*

III. Formant Amplitude Modification

One of the major tasks in modification of vocal effort was to be able to modify the amplitude of the formants. Amplitude of the formants cannot be directly modified form the LPC polynomial $A(z)$. To work around this problem, each formant is modeled as a two–pole digital resonator with a pair of complex conjugate poles located near the unit circle ($p_{1,2} = re^{\pm j\omega_0}$). For $r \approx 1$, the angle of the pole (ω_0) is related to the formant frequency (F_k) and the radius of the pole (r) is related to the formant bandwidth (B_k) as seen in the Equation 2, where ω_0 is the center frequency of the digital resonator.

The bandwidth of the resonator can be approximated to

$$B_{w_0} \approx 2(1 - r) \tag{2}$$

Given the formant (F_k), its bandwidth can be approximated by using relationship between the group delay and the bandwidths for single pole systems [6]

$$B_k \approx \frac{-1}{\pi}\left[ln\left(\frac{\partial \angle A(f)}{\partial f}\right) - ln\left(2\pi + \frac{\partial \angle A(f)}{\partial f}\right)\right]\Big|_{F_k} \tag{3}$$

where

$$\frac{\partial \angle A(f)}{\partial f} = Im\left\{\frac{\frac{\partial}{\partial f}A(f)}{A(f)}\right\} \tag{4}$$

$$\frac{\partial A(f)}{\partial f} = j2\pi \sum_{k=1}^{p} k a_k e^{-j2\pi fk} \tag{5}$$

$$A(f) = 1 - \sum_{k=1}^{p} a_k e^{-j2\pi fk} \tag{6}$$

Equation 6 is the LP spectrum and 5 is the derivative of $A(f)$

Using Equations 2 and 3 it is possible to estimate the pole location 'r' from the LPC polynomial without explicit root solving. The accuracy of the bandwidth estimation and thus 'r' estimation using 4, depends upon the accuracy of the formant estimate.

Give the location of the pole amplitude at center frequency ω_0 for the two–pole digital resonator using [7] is given by Equation 7.

$$|H(\omega_0)| = \frac{1}{(1-r)(\sqrt{1+r^2-2r\cos(2\omega_0)})} \tag{7}$$

As seen from 7 the amplitude of the formant depends on the angle and the radius of the pole. It is possible to modify the amplitude of the formant by changing one or both the parameters. Because we are interested in changing the amplitude of the formant without changing the formant frequency, the only parameter to change is the radius of the pole. Changing the radius will modify the bandwidth of that formant. The new radius ($\tilde{r}$) for ΔH change in the amplitude of the formant can be obtained from Equation 8

$$|H(\omega_0)| + \Delta H = \frac{1}{(1-\tilde{r})(\sqrt{1+\tilde{r}^2-2\tilde{r}\cos(2\omega_0)})} \tag{8}$$

The change in bandwidth due to the amplitudes modification can now be found using $\Delta B_k = 2(r - \tilde{r})$. LSP based formant bandwidth modification method as described in [8] was used to apply these bandwidth changes and get the new set of LPC coefficients.

An example of the algorithm operating on actual data can be seen in Figure 4 which shows LP spectrum

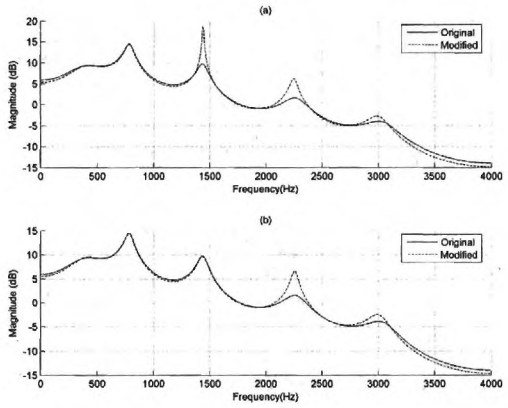


Fig. 4. *LP spectrum showing Formant Amplitude Modification*

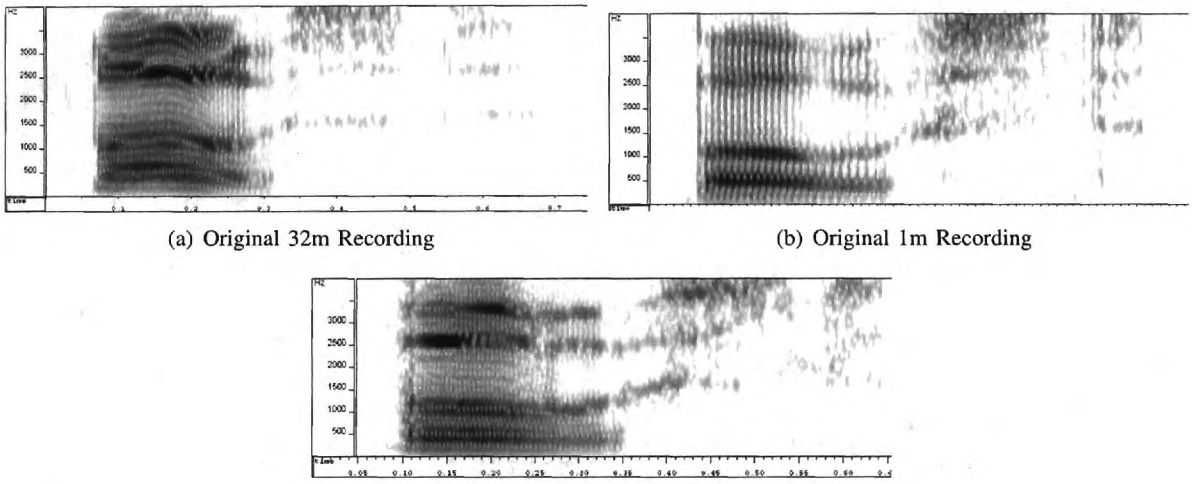
for original and modified polynomial. Modifications include the increase in amplitudes of second, third and the fourth formant by 8, 4 and 2 dB respectively as seen in (a) and in part (b) of the figure, amplitudes of third and fourth formants were modified by 4 and 2 dB respectively.

IV. Modification of Speech

As described in Section 2, modification of pitch, overall energy of the signal and multiple formant amplitudes (spectral tilt) are necessary to change the vocal effort to match a specified talker–to–listener distance. MELP vocoder framework was used to analyze the recorded speech, and also re–synthesize the speech with modified parameter set. The pitch, LPC coefficients and gain obtained during the analysis phase were modified in line with the requirement. Words are recorded at talker–to–listener distance of 1m were modified to match the vocal effort for talker–to–listener separations of 16 and 32 m. For male speakers, the pitch was increased by 30 Hz for 16 m separation and 65Hz for synthesizing 32 m separation. For female speakers, the pitch was increased by 20 and 40 Hz for 16 and 32 m separation respectively. Spectral tilt and gain were increased by 2/4 dB/Oct and 8/15 dB to synthesize vocal effort for 16 and 32 m respectively. Figure 5(c) shows the spectrogram of the synthesized speech. The vocal effort in 5(b), 1m talker–to–listener separation was modified to match the vocal effort of 5(a) 32 m talker–to–listener separation.

V. Perceptual Test and Results

To evaluate the performance of our algorithm two perceptual tests were performed on the data. In some initial informal tests, subjects were divided in two groups. Group 1 was presented with an audio at



(a) Original 32m Recording

(b) Original 1m Recording

(c) Synthesized 32m

Fig. 5. *Spectrograms for utterance of "mall"*

arbitrary talker–to–listener separations. Group 2 was first presented with an anchor, the audio recorded for talker–to–listener distance of 1m, henceforth referred as *anchor audio*, followed by an audio with arbitrary talker–to–listener separation(16 or 32m). The success rate for correct identification of separation was higher for Group 2. Based on these results the formal test included an anchor audio. This helped the test subjects to get acquaint themselves with the recording environment, and also provided a baseline to make decisions.

The 16 and 32m talker–to–listener separations were only considered in these tests. For the first test, the subjects were presented with an anchor audio followed by the audio clip of unknown talker–to–listener distance (16 or 32m). Two measures were collected as the part of this test. First, quality of the audio was assessed by having the test subject make a selection from 1 through 5 where input 1 meant "poor"audio quality and 5 meant "excellent". The second measure assesses the perceived talker–to–listener distance. The subject had to select from either 16m or 32m. Both the audio files were for the utterance of the same word by the same talker.

The second test played two audio clips for the subjects. Each clip had the same word uttered by same talker, but at different distances. The test subject was asked to identify the which clip represented a talker–to–listener distance of 32m. This test was aimed towards verifying whether the test subject can distinguish between the two listener talker distances.

Tests 1 and 2 both consisted of male and female original audio recordings at talker–to–listener distance of 1,16 and 32m, and male and female modified synthesized audio for talker–to–listener distance 16 and

32 m. The test subjects had no knowledge which audio clips(original/synthesized) were being played. Further, each test subject was given equal number of original and synthesized audio for both Test 1 and 2.

Twenty test subjects participated in the test. The results of Test 1 are documented in Tables I and II. Table I shows that quality of the audio on the scale of 1–5 as perceived by the listener. There is very little difference in the audio quality scores for the original recording and the synthesized audio, for both the male and female speakers. The modified and synthesized audio quality is as good as the original recordings.

Table II lists the percentage error in the perceived distance by the subject for both the male and female speakers. This score was also collected for original as well as synthesized audio. The error for male speakers is almost comparable suggesting the synthesis algorithm is performing well for male talker. The performance of the algorithm for female talkers is little better. Increase in the pitch is the one of the key parameter that influences the distance perception heavily. In the original recordings for some female talkers we did not observe a significant increase in pitch with distance, though there was a increase in formant amplitudes and overall energy of the signal, such utterances were ambiguous and were misclassified by the subjects and are the reason for high error rate. Contrary to this modified and synthesized sentences do have the required pitch variation and have lower probability of misclassification.

Test 2 also shows similar trend as Test 1. Original recording of 16m talker–to–listener separation is 22.91% of the time misclassified to be as be in 32m. This misclassification rate for modified and synthesized speech was found to be 23.95% .

TABLE I
Quality of Speech.

Audio	Male Talkers	Female Talkers
Original	3.90	3.83
Synthesized	3.5	3.63

TABLE II
Percentage Error in Perceived Distance.

Audio	Male Talkers	Female Talkers
Original	30.00%	34.09%
Synthesized	28.57%	21.42%

VI. Conclusions

An efficient algorithm for formant amplitude modification has been proposed. Acoustic parameters associated with change in vocal effort with distance from listener were identified. Further tests have confirmed that the variation of these parameters does indeed change the vocal effort and it is possible to modify an arbitrary recorded speech and to make it sound distant. The perceptual tests have also confirmed that the modification and synthesis algorithm produces intelligible speech with comparable quality to the original recordings that it is modifying.

References

[1] K. R. Bungart, D. S. Scott, "The effects of f0 manipulation on the perceived distance of speech," *Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 425–440, 2002.

[2] Brungart D. S. Kordik A. J. Das K. Shaw A. K., "The effects of production and presentation level on the auditory distance perception of speech," *Proceedings of the International Conference on Spoken Language Processing, Denver, Colorado*, pp. 1641–1644, 2001.

[3] A. Eriksson and H. Traunmüller, "Perception of vocal effort and distance from the speaker on the basis of vowel utterances," *Perception and Psychophysics*, vol. 64, no. 1, pp. 131–139, 2002.

[4] J. S. Lienard and M. G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *Journal of the Acoustical Society of America*, vol. 106, no. 1, pp. 411–422, 1999.

[5] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, 2002.

[6] R. W. Oppenheim, A. V. Schafer, *Discrete–Time Signal Processing*, Prentice–Hall, NJ, 1989.

[7] J.G. Proakis and D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, Prentice–Hall, 1996.

[8] M. W. Morris and M. A. Clements, "Modification of formants in the line spectrum domain," *IEEE Signal Processing Letter*, vol. 9, no. 1, pp. 19–21, 2002.